DOCUMENT RESUME

ED 210 309 TM 810 960

AUTHOR Powers, James E.
TITLE A Bayesian Method for Assigning Course Grades.
PUB DATE 81
NOTE 20p.; Paper presented at the Annual Meeting of the New England Educational Research Organization (Lenox, MA, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; *Bayesian Statistics; *Grading; Mathematical Formulas; Probability; *Student Evaluation

ABSTRACT

The use of Bayesian methodology to assign grades in classroom situations is presented. Assigning a grade is viewed from a criterion, as opposed to norm, referenced perspective. Criteria include mastery of some proportion, determined by the teacher, of the subject matter covered in a course. Different levels of mastery are deemed possible and, therefore, the assignment of different grades to different students may be appropriate. A prior probability distribution is assigned to the level of mastery, in terms of proportions, of each student at the outset of the course. After the first test is administered in the course, a likelihood distribution is obtained for each student. This distribution reflects the probability that the observed score would have been obtained, given each previously specified level of mastery. A posterior probability distribution on level of mastery is obtained for each student by combining the prior and likelihood distributions through the application of Bayes formula. For all subsequent tests, the posterior from the previous test is taken as the prior for the current test. A complete example illustrating the necessary calculations is included as well as a brief description of the Bayesian method.
(Author/DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED210309

A Bayesian Method for Assigning Course Grades

James E. Powers

State University of New York at Albany

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. E. Powers
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented at the 1981 meeting of the

New England Educational Research Organization

Lenox, Massachusetts

TM 810 960

A Bayesian Method for Assigning Course Grades

James E. Powers

State University of New York at Albany

At the conclusion of a course of instruction, an instructor must typically combine a number of measures of student achievement in some rational fashion in order to assign grades.

When these measures are test scores, a frequently recommended method for combination is the formation of composite scores. Typical methods for devising these composites, involving the calculation of standard, or at least semi-standard scores, only equate differential variability among the various measures (e.g., Payne, 1974). Thus these methods are only approximations, since in order to insure accurate weightings, covariance between measures must also be equated. The complexity of the derivation of composite score formulas which take into account both variance and covariance is prohibitive to their general classroom use. The mathematics necessary for these methods is given by Dunnette and Hoggatt (1957).

The purpose of this paper is to demonstrate a straightforward application of Bayes formula that can be used effectively, and relatively simply, in the classroom to aid in student evaluation. The suggested procedure obviates the need for the calculation of composite scores with all of their associated problems. Additionally, due to the unique capability of the Bayesian method to "carry" information from one experiment to another, it enables a teacher to keep a current estimate of a student's total achievement as a course progresses. Another advantage of the proposed method over more traditional methods is that it allows the flexibility of incorporating appropriate non-test information about students into the grading process in a quantitative manner.

The Bayesian Method

For the benefit of those readers who may not be familiar with the Bayesian approach to research the following brief, simplified discussion is presented.

Probably the most basic difference between the Bayesian and classical persuasions is how "probability" is defined. Within the Bayesian framework, probability may be operationally defined as simply the degree of belief an individual has that an event will occur. The classical viewpoint is that the probability that an event will occur is the proportion of times the event does occur in a large number of trials. Of the two definitions, the Bayesian is more general in that it allows an experimenter to either render a purely subjective judgment, or conduct a number of trials, to determine the probability in question. The important consideration is that all pertinent information is evaluated honestly. It is the element of subjectivity allowed by the Bayesian definition that has caused some controversy between the two viewpoints.

Beyond the difference in definition of probability, a Bayesian takes a somewhat different approach to experimentation than does a classicist. This approach basically involves evaluating as accurately as possible, subjectively if necessary one's belief about the parameter(s) in question at the present time, conducting an experiment to obtain new information and deciding how the original belief structure should be modified in the light of the new evidence.

To accomplish this, the following three steps are usually followed:

1. Specifing a prior distribution of belief about the parameter(s) in question.

2. Conducting an experimental investigation to obtain sample evidence about the parameter(s).

3. Combining the prior belief and sample evidence, via Bayes formula, to obtain a posterior distribution of belief about the parameter(s).

The prior distribution is a weight function that depicts the experimenter's belief about the unknown parameter prior to the collection of data. It theoretically represents everything the experimenter knows about the parameter at that time. To properly specify a prior distribution it is necessary that an experimenter evaluate relevant literature, personal experience, and any other source of appropriate information, and quantify this data as carefully as possible. Obviously, this frequently requires the use of subjective probability.

Quite often prior distributions are specified in terms of probabilities. It is conceptually convenient to think of a total probability of one to be assigned to the values of the parameter and then to decide upon the most accurate distribution. Actually, the prior must only show how many times as likely one value of the parameter is, compared to the other possible values.

Suppose, for example, that only two values of a parameter are possible and one desires to describe them as equally likely in the prior distribution. The assignment of probabilities of .5 and .5 will accomplish this purpose, The assignment of any two weights of equal magnitude, such as 11 and 11, will also do so, i.e., .5/.5 = 11/11 = 1.

Once the prior has been delineated an experiment is conducted to obtain data. The experimenter now calculates the likelihood of obtaining that data given each hypothesized value of the parameter.

Finally, the prior probabilities and the likelihoods are combined through Bayes formula to obtain a posterior distribution of belief. Now, assuming the prior distribution represented all previous knowledge, the posterior shows one what his beliefs should be about the parameter at this new point in time when <u>all</u> relevant information is considered. Any, and all, decisions about the unknown parameter are now made using the posterior probabilities.

Every Bayesian analysis, at some point, employs Bayes formula. The formula, for the sake of simplicity, may be written as:

$$(1) \quad P(\theta_i|X) = \frac{P(\theta_i)\ P(X|\theta_i)}{\sum_i P(\theta_i)\ P(X|\theta_i)}$$

$P(\theta_i)$ is the prior weight assigned to the i th value of the parameter.

$P(X|\theta_i)$ is the probability (likelihood) of observing the data X, given the i th value of the parameter.

$P(\theta_i|X)$ is the posterior probability of the i th value of the parameter, given the data.

The denominator on the right side of the expression is a scaling factor, simply a constant obtained by summing all possible numerators, that makes the posterior probabilities sum to unity for convenience.

An Example

To illustrate the use of the formula, as well as how the Bayesian method "carries" information from one experiment to another, consider the following ficticious example.

Suppose an instructor in an elementary statistics course presents the concept of standard scores during a lecture period. A student misses the lecture due to illness and the instructor spends a few minutes with him individually to explain the topic. Then, to determine whether or not the student understands the idea, the instructor asks three true-false questions. The questions are constructed to be quite simple if one understands the concept of standard scores.

Assume that the student gets all of the questions correct. The instructor must now ask whether the student understands the concept or simply guessed correctly.

Application of Bayes formula will help resolve the dilemma. The "parameter" in question is "understanding of the concept of standard scores." This is not

a parameter of the sort that one usually thinks of, but a "parameter" none the less. Suppose that in the instructor's opinion it is dichotomous, either it is understood or it is not. (One could consider "degrees of understanding." In that case the procedure would be somewhat different from the following illustration.)

Define the event E_1 to be that the student understands the concept after this first presentation and E_2 to be that he does not. The data, A, is that the student answered the three questions correctly. In this case Bayes formula is written as:

$$(2) \quad P(E_1|A) = \frac{P(E_1)\ P(A|E_1)}{\sum_i P(E_i)\ P(A|E_i)}$$

To make the application, it is necessary to obtain the prior probabilities and the likelihoods. Suppose that, having taught the course many times before, the instructor knows that one fourth of his students usually grasp the concept immediately and three fourths require additional explanation and/or practice problems before internalizing it. Having no additional information about this particular student, he assigns prior probabilities, $P(E_1)$ and $P(E_2)$, of 1/4 and 3/4 respectively.

Since each question has been constructed such that if one understands the concept he is certain to get it correct, the probability of one correct response, given understanding, is unity. Hence, the probability of three correct responses, given understanding, $P(A|E_1)$, is $1 \times 1 \times 1 = 1$.

If the concept is not understood, the student must guess and the probability of a correct response to one particular question is 1/2. Hence the probability of three correct, given no understanding, $P(A|E_2)$, is $1/2 \times 1/2 \times 1/2 = 1/8$.

Application of Bayes formula as given in equation (2) now gives:

$$P(E_1|A) = \frac{P(E_1)\ P(A|E_1)}{P(E_1)\ P(A|E_1) + P(E_2)\ P(A|E_2)}$$

$$= \frac{1/4\ (1)}{1/4\ (1) + 3/4\ (1/8)} = 8/11$$

$$\text{and } P(E_2|A) = \frac{P(E_2)\ P(A|E_2)}{P(E_1)\ P(A|E_1) + P(E_2)\ P(A|E_2)}$$

$$= \frac{3/4\ (1/8)}{1/4\ (1) + 3/4\ (1/8)} = 3/11$$

Since the posterior probabilities sum to unity, only one application of the formula was actually necessary. The second posterior probability could have been obtained by subtraction.

At this point, from the instructor's point of view, the probability that the concept is understood is 8/11. Whether or not this probability is convincing is a matter of individual choice. For the sake of the example let us assume that the instructor is not yet convinced and decides to "replicate" his experiment. To do this, he presents the student with two more questions constructed in the same fashion as the first three. The student gets both of these correct.

The instructor now reapplies Bayes formula with $P(A|E_1) = 1/4$ and $P(A|E_2) = 3/4$. The data, A, now becomes two correct responses, so that $P(A|E_1) = 1 \times 1 = 1$ and $P(A|E_2) = 1/2 \times 1/2 = 1/4$. Hence,

$$P(E_1|A) = \frac{1/4(1)}{1/4(1) + 3/4(1/4)} = 4/7$$

$$\text{and } P(E_2|A) = 3/7$$

At this point, the instructor does not know the true probability that the student understands the concept. The first experiment indicates that it is 8/11 and the second that it is 4/7 Since the outcomes of the two experiments are certainly not mutually exclusive and most probably not independent, the

probabilities are not simply additive or multiplative. Most readers would probably agree that the true probability is greater than 8/11 (though a surprising number would expect it to be between 4/7 and 8/11), but an exact figure is elusive.

In this simple example it is obvious that had all five questions been administered at the same time, the resulting experiment would have been equivalent to the two shorter ones. Applying Bayes formula to the larger experiment should then give the correct probability. Taking $P(E_1) = 1/4$ and $P(E_2) = 3/4$ as before, and $P(\ |E_1) = (1)^5 = 1$ and $P(A|E_2) = (1/2)^5 = 1/32$, we have:

$$P(E_1|A) = \frac{1/4(1)}{1/4(1) + 3/4(1/32)} = 32/35$$

$$\text{and } P(E_2|A) = 3/35$$

Then, if five trials had been conducted at the outset, beginning with the prior probabilities of 1/4 and 3/4 respectively, that the student did, or did not, understand the concept, the corresponding posterior probabilities of 32/35 and 3/35 would have been obtained.

Let us reconsider this example for a moment, recalling that a prior distribution should represent everything known about a parameter at that point in time. Thus, before administering the first three questions, the only information available to the instructor was that usually about 1/4 of his students understand the concept after one presentation and 3/4 do not. Therefore, those prior probabilities were reasonable. However, after the student had correctly responded to those questions, more information was available and new priors were in order for the second two questions. At that point in time, all of the information available about the student was contained in the posteriors derived from the first three questions and these were the most appropriate priors for the two question experiment.

If that had been done, then Bayes formula would have been employed with $P(E_1) = 8/11$, $P(E_2) = 3/11$, $P(A|E_1) = 1$ and $P(A|E_2) = 1/4$, with the result that:

$$P(E_1|A) = \frac{8/11(1)}{8/11(1) + 3/11(1/4)} = 32/35$$

$$\text{and } P(E_2|A) = 3/35$$

Thus the repeated application of Bayes formula to a series of test results is a convenient way to acquire cumulative information about students in a logical fashion. If done consistantly as a course progresses, a teacher will always have a current estimate of students total achievement at hand. It is only necessary that the original prior distributions be specified as accurately as possible and that each derived posterior be taken as the prior for the next examination.

General Use of the Method

In order to make the suggested application, let us conceptualize student achievement in a basically criterion referenced framework. The term "criterion referenced" is used here in a general sense. It simply means that a student's performance on a test will be compared to some criteria different from other students' performance as would be done in a norm referenced approach. Many authors present this basic definition (e.g. Glaser, 1963; Hills, 1976; Payne, 1974; Popham and Husek, 1969; Tyler, 1970), but then apply further restrictions to it. Quite frequently as soon as the term "criterion referenced" is used, the discussion immediately turns to mastery learning, where all, or most, students are expected to master all, or most, of the material, and where the usual criterion is a very detailed list of behavioral objectives.

In the present usage, we regard a test as a sample of questions taken from a large population of questions that could be asked about a particular body of subject matter. The criteria is mastery of some percentage, specified by the

teacher, of that subject matter. Further, different levels of mastery are deemed possible, indeed probable. In preparing items for such a test, the primary concern is that each question be representative of the population, i.e., content validity be high. The suggestions of Popham & Husek (1969) for the preparation of criterion, as opposed to norm, referenced tests should be followed.

For the purpose of this example, suppose that a teacher decides that mastery of 70% or more of the subject matter in a course warrants an "A", 60 to 79% a "B", 40 to 59% a "C", 20 to 39% a "D" and less than 20% an "F". At the end of the course, grades will be assigned depending upon the most likely level of mastery for each student.

To make the suggested Bayesian application, it is necessary for the instuctor to assign a prior distribution to the level of mastery for each student at the beginning of the course. After each quiz is administered, the posterior distribution obtained is then employed as the prior for the next quiz. Thus at the end of the course, the instructor will have a posterior distribution on the level of mastery of each student which contains all of the available information about that student and from which grades may be assigned.

To illustrate the method, we shall follow one student through a fictitious course where three quizzes were administered. Table 1 shows the mechanics of determining the posterior for this student after the first quiz. It is assumed that the quiz consisted of 20 questions, constructed to be independent, and that the student correctly answered 14.

Insert Table 1 about here

For this student, the individual prior probabilities assigned might have been selected by reasoning similar to that which follows. About 75% of the students who take the course master between 40 and 80% of the material. Therefore

.75 of the probability is assigned to this interval. The instructor has had no prior experience with this student and has no other information available. Thus, there is no reason to expect any particular level is more likely than any other, and the .75 is evenly distributed across the interval.

The remaining 25% of those who take the course master either 90 or more, or 30 or less, percent of the material, with slightly more at the upper, rather than the lower, extreme. Thus the remaining .25 is assigned to reflect that fact.

Notice that the levels of mastery above .95 and below .20 do not appear in the table. This means that the instructor has effectively assigned zero prior probability to these values. They are regarded as impossible. If they were regarded as possible, but not probable, they might have been included and assigned quite small weights.

Also notice that had the instructor had other information about the student it could have been incorporated in the prior. For example, if the student had been in a previous course with the instructor and performed quite well, a prior might have been assigned that favored the higher levels of mastery to a greater extent.

Since the questions on the test are constructed to be independent, they constitute a series of Bernoulli trials and the likelihoods may be obtained by evaluating the appropriate binominal distributions. The general form for binomials is given by:

$$(3) \quad P(X|N, \pi) = C_X^N \pi^X (1-\pi)^{N-X}$$

Thus the likelihood of a score of 14, if the level of mastery is .60 is:

$$P(X = 14|N = 20, \pi = .6) = C_{14}^{20} (.6)^{14} (.4)^6 = .124$$

If the level of mastery is .80, the likelihood is:

$$P(X = 14|N = 20, \pi = .8) = C_{14}^{20} (.8)^{14} (.2)^{6} = .109$$

and so forth.

Needless to say, if the calculations are to be done by hand, rather complete tables of binomials will eliminate considerable labor.

All of the numerators needed in the evaluations of Bayes formula are given in the next column of the table. These are simply the products of each prior probability times its respective likelihood, $P(\pi)P(X|\pi)$. Since the denominator of Bayes formula is simply the sum of all possible numerators, adding the entries in this column provides the scaling factor needed to make the posterior probabilities add to one.

Finally, the posterior probabilities are obtained by dividing each individual numerator by the sum of the numerators, e.g., .0009/.0711 = .013.

Table 2 shows the priors and posteriors for the remaining two tests in the course. Test 2 was taken to have N = 15 items and the students score was X = 9. For Test 3, N = 20 and X = 12.

Insert Table 2 about here

Now at the conclusion of the course when the instructor must make decisions about grades, the posterior probabilities from the last test are used. Assuming that the instructor quantified all available pertinent information as accurately as he could in the original prior distribution, and then employed successive applications of Bayes formula to accumulate additional information throughout the course, these final posterior probabilities should be an extremely good representation of the students' mastery, in terms of proportions, of the material.

The most appropriate grade for this student may now be determined by simple addition of the appropriate posterior probabilities. Table 3 summarizes this.

For this student, it would certainly appear that a B is the most reasonable grade. For other students, the situation might not be as clear. In these cases, the calculation of posterior odds ratios may be of assistance in making a decision. For example, the posterior odds that this student's mastery represents B, rather than A, performance are .856/.135 or 6.34 to 1. Similarly, the odds that mastery represents B, rather than C performance, are .856/.008 or 107 to 1 and those for B opposed to any other grade are .856/(.135 + .008) or 5.99 to 1.

Discussion

The use of Bayesian methods to evaluate student performance seems eminently reasonable to this author. This is in large part because the method allows one to attempt to explicitly quantify any subjective judgements that must be made through the specification of prior distributions. It is difficult to imagine a process for assigning grades that does not require some subjective judgments. However, the traditional procedures make no provision for dealing with them as thoroughly and openly as possible. Instead, an instructor is left to use informal, imprecise methods.

Prior distributions incorporating subjective judgments, though seen as distinctly advantageous by this author, are not seen so by some others. One objection, often stated, is that somehow one biases the results by using subjective judgments. My personal view is that, in the final analysis, results are probably less biased when one thinks carefully about these judgments and deals with them overtly, than when one pretends they don't exist. For those who do insist on maintaining "objectivity", however, the method should still be quite useful. If a rectangular (all levels of mastery assigned equal weights) prior distribution is employed initially, then all of the information accumulated comes from test scores and the final posterior should be "objective."

I do not recommend the use of a completely rectangular prior under any circumstances, even if one has very little information about a student. A totally rectangular prior is a plea of absolute ignorance. It says that "I have absolutely no information and therefore cannot even begin to differentially weight the various possible levels." A mastery level of zero must be treated as likely as one of 50%, and this is not credible regardless of how little information one has. The most abysmal lack of information imaginable seems to dictate at worst a "modified rectangular" prior. One similar to that used in the foregoing example; diffuse about the central values, but with smaller weights at the extremes.

It should also be noted that priors may be modified at any time between examinations. Thus if any new, pertinent, information is discovered about a student, it can be incorporated when appropriate. In the previous example, some modification of priors might have been desirable, not because of new information, but because finer distinctions between mastery levels could have been attained. Notice, in Table 2, that after the second test was given, the posterior probabilities are zero for the mastery levels .40 and less, and .90 and above. It certainly appears that this student is operating somewhere above .40 and below .90. Thus to continue to include values outside that range is a waste of effort. If, after the second test, the prior for Test 3 (posterior from Test 2) had been redefined to span only the levels of .45 to .85 in increments of .05, finer distinctions could have been made. Simple linear interpolation between the values in the posterior from Test 2 will give the appropriate prior for this new set of values.

One cannot pretend that there is no arithmetic necessary to use this method, but the labor required really does not seem substantially more than that needed to obtain more convential composite scores. Many, or most, of the usual methods

require the calculation of any, or all, of means, standard deviations, standard scores and composite score equations for all tests and/or students. The method suggested here does not require these things. Thus it appears that differences in labor are more in kind than in quantity.

A last word about calculations is that there are times when the normal approximation can be employed instead of the exact binomial distribution. This effects a substantial savings in labor. The usually accepted criteria for appropriateness of the approximation is that both $N\pi$ and $N(1-\pi)$ be no less than 5. So if a test had $N = 100$ items and levels of mastery were specified ranging from $\pi = .05$ to $\pi = .95$, the approximation could be used to obtain all likelihoods. Similarly, if $N = 50$ with $.10 \leq \pi \leq .90$, it could also be employed.

All things considered, it appears to this author that the method offers many advantages and is certainly as good as, and probably better than, most methods currently used to arrive at course grades.

Table 1

Calculation of a Posterior Probability Distribution on Level of Mastery for One Student

Proportion of Material Mastered	Prior Probability	Likelihood	Numerators	Posterior Probability
.95	.03	0	0	0
.90	.10	.009	.0009	.013
.80	.15	.109	.0164	.231
.70	.15	.192	.0288	.405
.60	.15	.124	.0186	.262
.50	.15	.037	.0056	.079
.40	.15	.005	.0008	.011
.30	.07	0	0	0
.20	.03	0	0	0
.10	.02	0	0	0
			.0711	

Table 2

Prior and Posterior Probability Distributions on Level of Mastery for One Student on Tests 2 and 3 in the Example

Proportion of Material Mastered	Posterior for Test 1, Prior for Test 2	Posterior for Test 2, Prior for Test 3	Posterior for Test 3
.95	0	0	0
.90	.013	.073	0
.80	.231	.437	.135
.70	.405	.397	.633
.60	.262	.089	.223
.50	.079	.005	.008
.40	.011	0	0
.30	0	0	0
.20	0	0	0
.10	0	0	0

Table 3

Final Probabilities that Level of Mastery Coincides with Various Grades for the Student in the Example

Grade	Level of Mastery	Probability
A	$\geq .80$	.135
B	.60-.79	.856
C	.40-.59	.008
D	.20-.39	0
E	<.20	0

REFERENCES

Dunnette, M. D., & Hoggatt, A. C. Deriving a composite from several measures of the same attribute. Educational and Psychological Measurement, 1957, 17, 423-434.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Hills, J. R. Measurement and evaluation in the classroom. Columbus, Ohio: Charles E. Merrill, 1976.

Payne, D. A. The assessment of learning: cognitive and affective. Lexington, Massachusetts: Heath, 1974.

Popham, W. J., & Husek, T. R. Implications of criterion referenced measurement. Journal of Educational Measurement, 1969, 6 (1), 1-9.

Tyler, R. W. Testing for accountability. The Nation's Schools, 1970, 83 (6), 37-39.